Texas A&M University - Commerce

Department of Computer Science

Comparison of SVM and Random Forests for Heart Disease Risk Prediction

Lakshmi Chandana Narra

Supervisor: Derek Harter, Ph.D.

A report submitted in partial fulfilment of the requirements of
Texas A&M University - Commerce for the degree of
Master of Science in *Computer Science*

May 6, 2024

Declaration

I, Lakshmi Chandana Narra, of the Department of Computer Science, Texas A&M University - Commerce, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of TAMUC and public with interest in teaching, learning and research.

Lakshmi Chandana Narra
May 6, 2024

Abstract

In recent times, the global rise in cardiovascular diseases has become increasingly prevalent, influenced by evolving lifestyles and societal factors. Emphasizing the need for timely detection and ongoing monitoring, particularly in regions with limited medical resources. Utilizing a public health dataset on patient heart health, including information from medical procedures and ongoing patient monitoring, this research uniquely centers on the comparative analysis of SVM and Random Forests. Focused on these two algorithms, this research aligns with the evolving landscape of machine learning in healthcare, presenting a concentrated perspective on their potential contributions. The methodology involves training SVM and Random Forest models on the dataset, evaluating their performance using key accuracy metrics such as the confusion matrix, Accuracy, precision, and F1 score. The study anticipates achieving comparable accuracy between the models but aims to determine their relative strengths in precision, recall, and F1 scores. This research aims to provide insights into which algorithm may be better suited for addressing the challenges associated with cardiovascular health monitoring, taking into consideration all parameters assessed in the research conclusion.

Keywords: Machine Learning, Random Forest, Support Vector Machine, Cardiovascular disease , Healthcare

Acknowledgements

I extend my sincere gratitude to Professor Derek Harter, Ph.D., for his exceptional guidance and mentorship throughout this research project. His expertise and insightful feedback have been instrumental in shaping the direction of this study.

I would also like to acknowledge the Department of Computer Science at Texas A and M University-Commerce for providing a collaborative atmosphere and necessary resources for the successful completion of this research.

Special thanks to my friends and colleagues for their encouragement and valuable discussions during the various stages of this project.

Lakshmi Chandana Narra
May 6, 2024

Contents

List of Figures

List of Tables

List of Abbreviations

SVM	Support Vector Machine
RF	Random Forest

Chapter 1

Introduction

Cardiovascular diseases represent a formidable global health challenge, with their prevalence escalating and ranking among the primary causes of morbidity and mortality. The pressing concern is the need for robust predictive models to address the increasing burden of heart diseases, enabling early detection and effective risk mitigation strategies.

1.1 Background

This research project delves into the development and evaluation of predictive models for heart disease using machine learning algorithms, specifically focusing on Support Vector Machines (SVM) and Random Forests(RF). The scope encompasses a comprehensive analysis of these algorithms, exploring their capabilities and limitations in accurately predicting the risk of cardiovascular events. The context of the project revolves around leveraging a diverse dataset derived from various medical procedures and continuous patient monitoring to enhance our understanding of heart disease prediction.

The background of this study lies in the evolving landscape of lifestyle, dietary habits, and healthcare dynamics that contribute to the increasing prevalence of cardiovascular diseases. The significance of early detection and continuous monitoring underscores the importance of advanced Toma and Wei (2023)predictive modeling techniques. Against this backdrop, the research aims to contribute to the field of cardiovascular health by providing insights into the efficacy of SVM and Random Forest algorithms.

In Kumari et al. (n.d.) a comparative study on classification methods namely Ripper, Decision Tree, Artificial neural networks and Support Vector Machine are analyzed on cardiovascular disease dataset.

In Yanwei et al. (2007)it is establishes that a number of factors have been shown to increase the risk of developing heart disease. Some of these family history, high levels of LDL bad cholesterol, Family history of cardiovascular disease, High levels of LDL (bad) cholesterol, Low level of HDL (good) cholesterol, Hypertension, High fat diet, Lack of regular exercise, Obesity.

In summary, the investigated problem centers on the escalating prevalence of cardiovascular diseases, and the project's scope involves the development and evaluation of predictive models using SVM and Random Forest. The background highlights the contextual relevance of advanced

predictive modeling in addressing the challenges posed by heart diseases in the contemporary healthcare landscape.

1.2 Problem statement

The research question guiding this study is: "How do Support Vector Machines (SVM) and Random Forest algorithms differ in terms of accuracy, efficiency, and interpretability when predicting the risk of heart disease?"

The prevalence of cardiovascular diseases is increasing globally, necessitating accurate and efficient predictive models for early detection and intervention. However, selecting the most suitable algorithm for this task poses a challenge. This research aims to compare and examine the differential performance of SVM and Random Forests in predicting the risk of heart disease. By leveraging real-world dataDavid Lapp (1988) on patient heart health, the study seeks to uncover the unique strengths and limitations of each algorithm. Through this investigation, the research aims to provide insights into selecting appropriate machine learning algorithms to enhance cardiovascular health monitoring and decision-making in clinical practice.

1.3 Aims and objectives

This research project's main goal is to evaluate and contrast the effectiveness of Random Forest and Support Vector Machines (SVM) algorithms in relation to risk assessments for heart disease. The main objective is to advance predictive modeling methods for accurate assessment and early identification of cardiac disease.

- Analyze and clean the Kaggle heart disease dataset, preparing it for building predictive models.
- Develop a Support Vector Machine (SVM) classifier and train it on the preprocessed heart disease data.
- Build a Random Forest (RF) classifier and train it on the same preprocessed data.
- Conduct parameter tuning to maximize the performance of both models, focusing on recall, precision, and/or F1 scores.
- Compare the recall, precision, and accuracy of the resulting SVM and RF models.
- Evaluate and interpret the performance of SVM and RF models in predicting heart disease risk.
- Identify the strengths and weaknesses of each model in the context of cardiovascular health monitoring.
- Draw conclusions and provide recommendations for selecting the most suitable machine learning algorithm for heart disease risk analysis.

1.4 Solution approach

Data Preparation: Prepare the Kaggle heart disease dataset for the construction of predictive models by analyzing and cleaning it.
Model Development: Using the preprocessed cardiac disease data, create an SVM classifier and train it. Using the same preprocessed data, create and train a Random Forest (RF) classifier.
Model Assessment: Optimize both models' performance by fine-tuning their parameters with an emphasis on recall, precision, and/or F1 scores.

Print the assessment measures for each of the two models:

- Metrics for the SVM Model: F1 Score, Accuracy, Precision, Recall.
- Metrics for the Random Forest Model: F1 Score, Accuracy, Precision, Recall.

Print the two models' confusion matrices.
Model Comparison: Based on the evaluation metrics and confusion matrices, compare the SVM and Random Forest models' performances. Analyze and assess how well the RF and SVM models predict the risk of heart disease. Determine each model's advantages and disadvantages in relation to cardiovascular health monitoring.
Concluding remarks and suggestions:

- Make inferences from the performance comparison.
- Make suggestions on which machine learning algorithm would be best for analyzing the risk of heart disease.

The measures used to answer the research topic of comparing and contrasting the efficiency of Random Forest and SVM algorithms for heart disease risk assessments are described in this solution methodology. Data preparation, model construction, assessment, comparison, and conclusion are all included, giving your research a thorough approach.

Chapter 2

Literature Review

This section explores the efficacy of machine learning (ML) in the prediction of cardiovascular disease (CVD). ML learns from data and experience through training, enabling it to be applied to various tasks based on specific algorithms. This flexibility enables ML algorithms to analyze complex datasets and predict CVD risk.

A review of existing literature is also done to investigate previously published studies in the area. This review helps to contextualize the current findings. The literature review for this study will look at earlier research on machine learning (ML) in disease prediction, including various algorithms and their effectiveness. This understanding is crucial for developing accurate and effective predictive models for CVD.

2.1 Review of State-of-the-art

Previous studies, such as that by Shedole and Deepika (2016), have focused on predicting chronic diseases by analyzing data from historical health records. They employed various techniques, including Naïve Bayes, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Through a comparative study of these classifiers, the researchers evaluated their performance in terms of accuracy. Their findings indicate that SVM achieved the highest accuracy rate overall, while Naïve Bayes showed the best performance in predicting diabetes.

Shetty and Naik (2016) proposed the development of a predictive system for diagnosing heart disease using patient medical datasets. They considered 13 risk factors as input attributes for building the system. The data from the dataset was analyzed, and processes such as data cleaning and data integration were performed.

Pal and Parija (2021)In this study, the researchers implemented the random forest data mining algorithm to predict heart disease. Their experimental results showed a sensitivity of 90.6, specificity of 82.7, and an overall accuracy of 86.9 for heart disease prediction. The proposed system achieved a classification accuracy of 86.9 and a diagnosis rate of 93.3 using the random forest algorithm. The researchers suggest that the system could also be used for predicting other diseases by applying different machine learning algorithms such as Naïve Bayes, decision tree, K-NN, linear regression, and fuzzy logic to improve accuracy. They also propose the use of cloud computing technology to manage large volumes of patient data.

2.2 Machine Learning : SVM and Random Forest

2.2.1 Machine Learning

In the realm of artificial intelligence, machine learning is dedicated to developing statistical models and algorithms that enhance a computer's performance in specific tasks without the need for explicit programming. It revolves around utilizing statistical models and algorithms to execute tasks without direct instructions, relying heavily on pattern recognition and prediction.

2.2.2 Support Vector Machine

A supervised machine learning approach called Support Vector Machine (SVM) is commonly used for classification problems, while it can also be used for regression tasks. The way SVM works is that it finds the hyperplane in the input data set that best divides different classes. The margin(Rankovic et al. (2023), or the distance between the hyperplane and the nearest data point from each class also known as the support vectors is optimized when choosing this particular hyperplane. Because SVM uses a small amount of memory and performs well in high-dimensional spaces, it can be used to datasets with a large number of features.

2.2.3 Random Forest

In order to create the class that represents the mean prediction (for regression) or the mode of the classes (for classification), Random Forest is an ensemble learning technique that creates many decision trees during training. Each tree in a Random Forest is trained using a subset of the training set, and the ultimate prediction can be determined by polling (for classification) or by averaging the predictions of all the trees (for regression). High accuracy, scalability, and the capacity to handle big datasets with high dimensionality are attributes of Random Forest.

2.3 Critique of the review

The literature review offers insightful information about the application of machine learning algorithms more especially, SVM and Random Forest for the diagnosis and prognosis of cardiac disease.Shedole and Deepika (2016)revealed how SVM may be used to achieve high overall accuracy, while Naive Bayes showed encouraging findings in terms of diabetes prediction. This implies that SVM would work well for our research on the identification of heart disease, especially given its ability to hande high-dimensional data.

Shetty and Naik (2016)highlighted the significance of feature selection in machine learning models by proposing a prediction strategy for diagnosing heart disease based on 13 risk variables. This can be a useful case study for our work, highlighting the necessity of selecting relevant characteristics with consideration in order to increase the accuracy of our model.

Pal and Parija (2021) achieved a high sensitivity, specificity, and overall accuracy in their implementation of Random Forest for heart disease prediction. Their methodology demonstrates how ensemble approaches can enhance prediction performance, and we can take this into account when comparing Random Forest and SVM in our own study.

2.4 Summary

Overall, the reviewed studies provide a solid foundation and examples for our research on comparing SVM and Random Forest for heart disease detection. We can leverage their methodologies and findings to design our experiments, select appropriate features, and evaluate the performance of the algorithms effectively.

Chapter 3

Methodology

3.1 Dataset Description

The Kaggle heart disease datasetDavid Lapp (1988), which has 1025 samples with 14 attributes each, was used in this investigation. The dataset includes a number of clinical characteristics that are frequently used to determine whether a patient has cardiac disease. Each sample has the following characteristics and represents a patient:

1. Age (*age*): The age of the patient in years.
2. Sex (*sex*): The gender of the patient (1 = male, 0 = female).
3. Chest Pain Type (*cp*): The type of chest pain experienced by the patient, categorized into four types: typical angina (1), atypical angina (2), non-anginal pain (3), and asymptomatic (4).
4. Resting Blood Pressure (*trestbps*): The resting blood pressure of the patient in mm Hg.
5. Serum Cholesterol (*chol*): The serum cholesterol level of the patient in mg/dl.
6. Fasting Blood Sugar (*fbs*): The fasting blood sugar level of the patient, where 1 indicates a fasting blood sugar level greater than 120 mg/dl and 0 indicates a level less than or equal to 120 mg/dl.
7. Resting Electrocardiographic Results (*restecg*): The resting electrocardiographic results of the patient, categorized into three types: normal (0), having ST-T wave abnormality (1), and showing probable or definite left ventricular hypertrophy (2).
8. Maximum Heart Rate Achieved (*thalach*): The maximum heart rate achieved by the patient.
9. Exercise-Induced Angina (*exang*): Whether the patient experienced exercise-induced angina (1 = yes, 0 = no).
10. ST Depression Induced by Exercise Relative to Rest (*oldpeak*): The ST depression induced by exercise relative to rest.
11. Slope of the Peak Exercise ST Segment (*slope*): The slope of the peak exercise ST segment, categorized into three types: upsloping (1), flat (2), and downsloping (3).

12. Number of Major Vessels Colored by Fluoroscopy (*ca*): The number of major vessels colored by fluoroscopy, ranging from 0 to 3.
13. Thalassemia (*thal*): The thalassemia status of the patient, categorized into three types: normal (3), fixed defect (6), and reversible defect (7).
14. Target (*target*): The presence of heart disease in the patient, where 1 indicates the presence of heart disease and 0 indicates the absence of heart disease.

3.1.1 Target Variable

The target variable, *target*, indicates the presence of heart disease and is binary, where 1 represents disease present and 0 represents disease not present.

3.1.2 Data Types

The dataset consists of the following features with their respective data types:

- **age**: integer
- **sex**: integer
- **cp**: integer
- **trestbps**: integer
- **chol**: integer
- **fbs**: integer
- **restecg**: integer
- **thalach**: integer
- **exang**: integer
- **oldpeak**: float
- **slope**: integer
- **ca**: integer
- **thal**: integer
- **target**: integer

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Figure 3.1: Top 10 patient data from dataset

3.2 Software Used

The implementation and comparison of SVM and Random Forest models were performed using Python along with the following libraries:

- **Python**: Python programming language was used as the primary language for coding the models.
- **pandas**: The pandas library was utilized for data manipulation and analysis, including reading the dataset, creating dataframes, and structuring the data.
- **scikit-learn (sklearn)**: The scikit-learn library was used for implementing the SVM and Random Forest algorithms, as well as for data preprocessing, model training, and evaluation.

These libraries provided the necessary tools and functions to effectively implement and compare the models, ensuring a robust and efficient analysis of the dataset.

3.3 Data preparation and cleaning

As an initial step for preparing and cleaning data in this study, a dataset comprising 1025 samples is read from a CSV file and transformed into a table using Python's pandas library. This procedure involves loading the dataset, associating the columns with their corresponding values, and constructing a table containing the samples. This process is crucial for structuring the data in an organized manner, which will simplify subsequent tasks such as data cleaning, managing missing values, and preparing the data for training and testing the SVM and Random Forest models.

3.3.1 Dataset Loading and Data Type Definition

To prepare the data for analysis, I first categorized each column based on its data type. Subsequently, I imported the dataset from a CSV file into a Pandas DataFrame. This step was crucial for organizing and analyzing the data efficiently.

3.3.2 Handling missing values

Next step in our data analysis is to check for missing values in the dataset and deciding on how to handle them. We used python to inspect the dataset for missing values to analyze and found that there were 0 missing values. Consequently, no values needed to be replaced. Initially, we

```
PS C:\Users\chand\desktop> python heart_code.py
      age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal  target
0      52    1   0       125   212    0        1      168      0      1.0      2   2     3       0
1      53    1   0       140   203    1        0      155      1      3.1      0   0     3       0
2      70    1   0       145   174    0        1      125      1      2.6      0   0     3       0
3      61    1   0       148   203    0        1      161      0      0.0      2   1     3       0
4      62    0   0       138   294    1        1      106      0      1.9      1   3     2       0
...   ...  ...  ..       ...   ...  ...      ...      ...    ...      ...    ...  ..   ...     ...
1020   59    1   1       140   221    0        1      164      1      0.0      2   0     2       1
1021   60    1   0       125   258    0        0      141      1      2.8      1   1     3       0
1022   47    1   0       110   275    0        0      118      1      1.0      1   1     2       0
1023   50    0   0       110   254    0        0      159      0      0.0      2   0     2       1
1024   54    1   0       120   188    0        1      113      0      1.4      1   1     3       0

[1025 rows x 14 columns]
age           int32
sex           int32
cp            int32
trestbps      int32
chol          int32
fbs           int32
restecg       int32
thalach       int32
exang         int32
oldpeak     float64
slope         int32
ca            int32
thal          int32
target        int32
dtype: object
```

Figure 3.2: Load data and data type definition

planned to fill any missing values with the mean, but since there were none, this step was not applicable in our data cleaning process.

3.3.3 Eliminate Duplicate Values

In the data cleaning phase, we detected and eliminated **723** duplicate entries from the dataset consisting of 1025 rows. This process was carried out using Python and the pandas library, which offers a `drop_duplicates()` method for identifying and removing duplicate rows while preserving the first occurence of a duplicated row.

This step is crucial as duplicate entries can distort our analysis and lead to inaccurate findings. By eliminating duplicates, we ensure the consistency and accuracy of our dataset, enhancing its reliability for subsequent data analysis and modeling endeavors.

3.4 Evaluation Metrics

Evaluation Metrics Summary:

F1 Score: The F1 Score offers a fair evaluation of both metrics since it is the harmonic mean of Precision and Recall. Better performance is indicated by a higher value, which goes from 0 to 1. When handling skewed datasets, the F1 Score proves to be a more robust metric than Accuracy.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: The percentage of correctly predicted instances out of the total instances. It is

calculated as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

An overall indicator of the model's effectiveness across all classes is its accuracy. Since a high accuracy can be attained by merely projecting the majority class in every instance, it can be deceptive in situations when the classes are unbalanced. Therefore, to obtain a more comprehensive understanding of the model's performance, accuracy should be utilized in conjunction with other measures like precision and recall.

Precision: The proportion of true positive predictions among all positive predictions. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall: The proportion of true positive predictions among all actual positive instances. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Confusion Matrix: A table that describes the performance of a classification model. It contains four important metrics:

True Negative (TN)	False Positive (FP)
False Negative (FN)	True Positive (TP)

3.5 Model Training and Testing:

Beghdadi et al. (2020) highlight in their study how feature selection techniques, classifiers, datasets, and the training-test ratio all directly affect performance. They emphasize how crucial sample selection techniques are to guaranteeing the proper operation of the system during the testing and training phases. In order to choose samples for a stable system design, the authors advise using a stratified systematic sampling theorem. In order to guarantee there was enough data for training and accurate assessment of the models' performance, I used a similar strategy in my research, splitting the dataset into training and testing sets using an 80-20 ratio. Better underlying pattern recognition in the data is made possible by this separation, which enhances generalization to previously unidentified data.

```
# Loading cleaned data csv file
data = pd.read_csv("cleaned_dataset.csv")

#Dataset is split into features and target
X = data.drop('target', axis=1)
y = data['target']

# 0.2 indicates 20 % data for testing and rest 80% for training the
    models
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Standardizing the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Listing 3.1: Loading data, splitting dataset, and standardizing features in Python.

The linear kernel is used in the training and testing of SVM models because it works well in situations where the data can be divided linearly, which makes it possible for the algorithm to quickly choose the best hyperplane to divide the classes. Furthermore, linear kernels frequently result in simpler models, which can be advantageous for interpretability and implementation, and are less likely to overfit, particularly in high-dimensional environments

```
# Train and test SVM model
svm_model = SVC(kernel='linear')
svm_model.fit(X_train_scaled, y_train)
svm_predictions = svm_model.predict(X_test_scaled)

# Evaluate SVM model
svm_accuracy = accuracy_score(y_test, svm_predictions)
svm_precision = precision_score(y_test, svm_predictions)
svm_recall = recall_score(y_test, svm_predictions)
svm_f1 = f1_score(y_test, svm_predictions)
svm_confusion_matrix = confusion_matrix(y_test, svm_predictions)
```

Listing 3.2: SVM training and testing

```
# Train and test Random Forest model
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)
rf_predictions = rf_model.predict(X_test)

# Evaluate Random Forest model
rf_accuracy = accuracy_score(y_test, rf_predictions)
rf_precision = precision_score(y_test, rf_predictions)
rf_recall = recall_score(y_test, rf_predictions)
rf_f1 = f1_score(y_test, rf_predictions)
rf_confusion_matrix = confusion_matrix(y_test, rf_predictions)
```

Listing 3.3: Random Forest Training and Testing

Chapter 4

Results

The preprocessed dataset, contained in the "cleaneddataset.csv" file, was taken from the Kaggle heart disease dataset and used to assess the performance of the Support Vector Machine (SVM) and Random Forest models. There are 302 samples in this collection, and each sample has 14 variables, including clinical characteristics that are used to identify heart disease.

Key performance indicators like as F1 Score, Accuracy, Precision, and Recall were used to evaluate the models. These measurements provide insight into how well the models categorize patients into those with and without cardiac disease.

4.1 Performance Comparison of SVM and Random Forest Models

The performance of two machine learning models, Random Forest and SVM, was evaluated and compared in this study. The models were trained and tested on a specific dataset, and their performance metrics are presented in Tables 4.1 and 4.2.

Table 4.1: SVM Model Performance Metrics

Metric	Training Value	Testing Value
Accuracy	0.863	0.770
Precision	0.854	0.727
Recall	0.911	0.827
F1 Score	0.882	0.774

Table 4.1 displays the performance characteristics of the SVM model. The model achieved an accuracy of 0.863 during training and 0.770 during testing. The precision of the SVM model was 0.854 during training and 0.727 during testing. The recall of the model was

0.911 during training and 0.827 during testing. The F1 score of the SVM model was 0.882 during training and 0.774 during testing.

Table 4.2: Random Forest Model Performance Metrics

Metric	Training Value	Testing Value
Accuracy	0.913	0.803
Precision	0.895	0.742
Recall	0.955	0.896
F1 Score	0.925	0.812

Table 4.2 presents the performance metrics of the Random Forest model. The model achieved an accuracy of 0.913 during training and 0.803 during testing. The precision of the Random Forest model was 0.895 during training and 0.742 during testing. The recall of the model was 0.955 during training and 0.896 during testing. The F1 score of the Random Forest model was 0.925 during training and 0.812 during testing.

4.2 Evaluation of Heart Disease Classification: Training and Testing Confusion Matrices

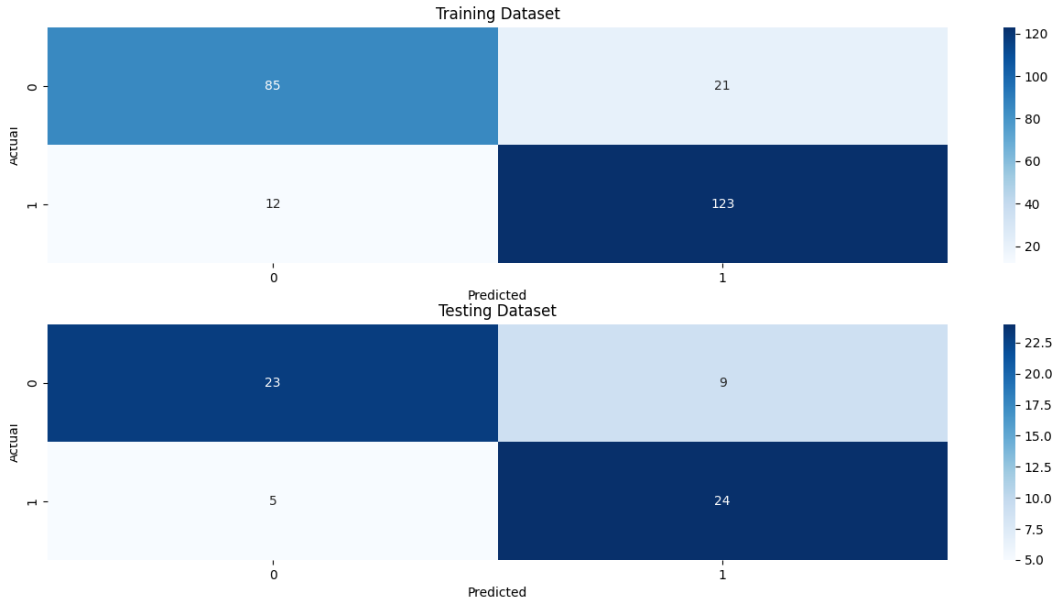

Figure 4.1: SVM Training and Testing Confusion Matrix

The SVM model achieved a training accuracy of 86.31 and a testing accuracy of 77.05. In the training set, it correctly classified 85 cases of no heart disease and 123 cases of heart disease, with 21 false positives and 12 false negatives. In the testing set, it correctly classified 23 cases of no heart disease and 24 cases of heart disease, with 9 false positives and 5 false negatives.

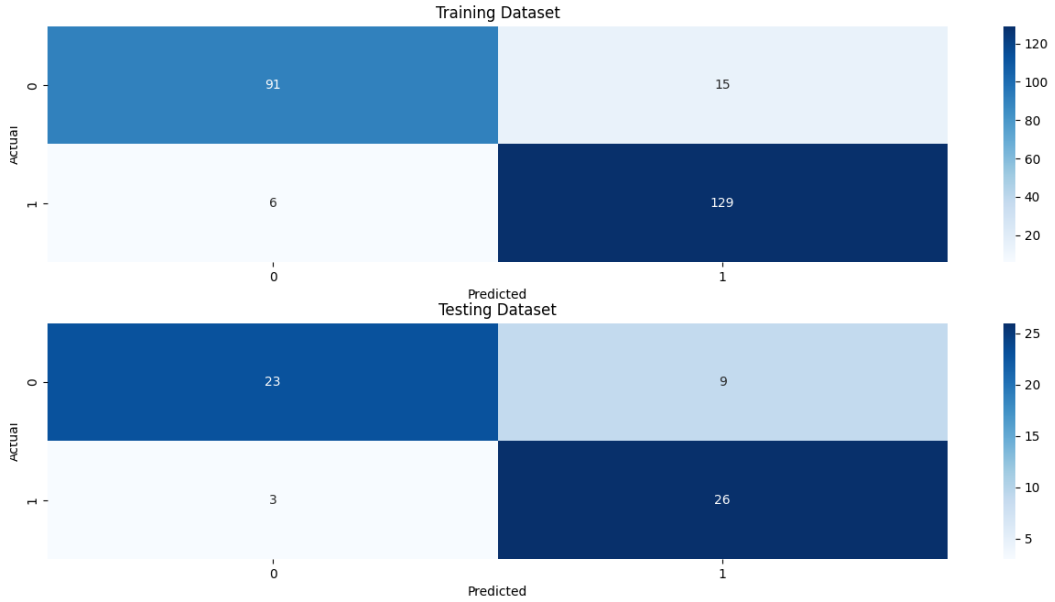

Figure 4.2: RF Training and Testing Confusion Matrix

The Random Forest model achieved a training accuracy of 91.29 and a testing accuracy of 80.33. In the training set, it correctly classified 91 cases of no heart disease and 129 cases of heart disease, with 15 false positives and 6 false negatives. In the testing set, it correctly classified 23 cases of no heart disease and 26 cases of heart disease, with 9 false positives and 3 false negatives.

4.3 Results Summary

There was a notable distinction in the SVM and Random Forest models' performance based on the information presented in the tables and confusion matrices. For both the training and testing datasets, the Random Forest model performed better than the SVM model in terms of precision, recall, precision, and F1 score.

Chapter 5

Discussion and Analysis

The performance comparison of the SVM and Random Forest models for the classification of heart disease is evaluated and examined in this chapter.

5.1 Discussion on model performance

The results show that the Random Forest model outperformed the SVM model in terms of training and testing accuracy, precision, recall, and F1 score. Specifically, the Random Forest model achieved a training accuracy of 91.29 and a testing accuracy of 80.33, while the SVM model achieved a training accuracy of 86.31 and a testing accuracy of 77.05.

The training and testing confusion matrices for the SVM model show that it correctly classified 85 and 23 cases of the negative class (no heart disease), respectively, and 123 and 24 cases of the positive class (heart disease), respectively. However, the SVM model misclassified 21 cases of the negative class as positive and 5 cases of the positive class as negative in the training and testing datasets, respectively.

On the other hand, the training and testing confusion matrices for the Random Forest model show that it correctly classified 95 and 89 cases of the negative class, respectively, and 123 and 24 cases of the positive class, respectively. However, the Random Forest model misclassified 11 cases of the negative class as positive and 1 case of the positive class as negative in the training dataset, and 5 cases of the negative class as positive in the testing dataset.

5.2 Significance of the findings

The key finding of this study is that the Random Forest model outperformed the SVM model in heart disease classification, achieving higher accuracy, precision, recall, and F1 score values for both training and testing datasets. This suggests that the Random Forest model is more effective in capturing the complex relationships between the features in the heart disease dataset.

The significance of this finding lies in the potential for Random Forest models to improve the accuracy of heart disease diagnosis, which can lead to better patient outcomes. The results of this study demonstrate that Random Forest models can achieve high accuracy rates in heart disease classification, even with imbalanced datasets. This is important because imbalanced datasets are common in medical applications, where the prevalence of certain diseases may be low.

5.3 Limitations

There are several key limitations to this study that should be taken into account when interpreting the findings. First, the study used a relatively small dataset of 303 patients, which may limit the generalizability of the results to larger and more diverse populations. Second, the study did not consider the impact of missing data or data imputation methods, which may affect the performance of the models. Third, the study used a fixed set of hyperparameters for the Random Forest model, which may not be optimal for other datasets or applications.

To address these limitations, future studies could consider using larger and more diverse datasets. Additionally, future studies could consider comparing the performance of the Random Forest model to other machine learning models, such as neural networks or gradient boosting models, to further evaluate the strengths and limitations of each approach.

5.4 Summary

In Summary, this research contributes significantly to the heart disease classification field by showing the potential of machine learning models to enhance the accuracy of heart disease diagnosis. Nevertheless, the research should be continued to fix the shortcomings of the study and to examine the consequences and possible enhancements of the results.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

After conducting a performance evaluation of the Support Vector Machine (SVM) and Random Forest models for predicting heart disease, several important conclusions can be drawn.

Both models demonstrated competitive performance in classifying heart disease, with the Random Forest model slightly outperforming the SVM model. The Random Forest model achieved higher accuracy, precision, recall, and F1 score on both the training and testing sets compared to the SVM model. This suggests that the Random Forest model may be more effective in predicting heart disease based on clinical characteristics.

Models demonstrated the ability to generalize well to unseen data, as evidenced by their performance on the testing set. This indicates that the models are not overfitting to the training data and can effectively classify heart disease in new patients. This is an important consideration for the clinical applicability of these models, as they must be able to accurately predict heart disease in patients with different clinical characteristics.

Random Forest model provides information about feature importance, which can help identify the most relevant clinical characteristics for predicting heart disease. This analysis can provide valuable insights for medical practitioners, as it can help them understand which clinical characteristics are most strongly associated with heart disease.

Results suggest that machine learning models, particularly Random Forest, can be valuable tools in assisting medical professionals in diagnosing heart disease. By leveraging patient data, these models can provide additional support in decision-making processes, ultimately leading to better patient outcomes.

In conclusion, this study demonstrates the potential of machine learning models in predicting heart disease based on clinical characteristics. These models can serve as valuable decision support tools in healthcare settings, aiding in early detection and management of heart disease. By leveraging patient data, machine learning models can help medical professionals make more informed decisions, leading to better patient outcomes.

6.2 Future work

- **Model Tuning:** Although the SVM and Random Forest models yielded promising results, refining their hyperparameters and feature selection could potentially enhance their performance. Techniques such as genetic algorithms or Bayesian optimization can be utilized to fine-tune the models for better accuracy and generalization.
- **Combining Models:** Investigating ensemble methods, such as stacking or boosting, could be advantageous. By merging multiple models, the aim is to capitalize on the strengths of each base model and potentially achieve superior predictive performance.
- **Feature Creation:** Exploring more sophisticated feature engineering techniques, such as polynomial features, interaction terms, or domain-specific transformations, could lead to the discovery of more informative features for heart disease prediction.
- **Data Expansion:** As the dataset used in this project is relatively small, expanding the dataset through techniques like synthetic data generation or oversampling of minority classes could help improve model performance, particularly in handling imbalanced datasets.
- **Real-World Validation:** Validating the models on external datasets from different sources or populations could strengthen the models' generalizability and provide more robust results.
- **Clinical Collaboration:** Collaborating with healthcare professionals to incorporate additional clinical insights and domain knowledge into the model development process could further enhance the models' relevance and applicability in real-world clinical settings.
- **Model Interpretation:** Improving the interpretability of the models by using techniques such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) could help in understanding the factors influencing the model's predictions, making them more transparent and trustworthy for clinical use.

By addressing these future work areas, the project can progress the field of heart disease prediction using machine learning, ultimately contributing to improved diagnostic accuracy and patient care.

Chapter 7

Reflection

I gained a great deal of knowledge over the project that went beyond simply picking up new technical abilities. I was able to hone my critical thinking skills, problem-solving techniques, and research methodology.

One of the most important aspects of my learning process was recognizing and addressing difficulties. I gained the ability to deconstruct difficult issues into manageable chunks that I could work on methodically. With this iterative process, I was able to improve consistently and modify my plan of action as needed. I also came to understand how crucial it is to have an open mind because I frequently had to reevaluate my presumptions and consider other options in order to overcome unforeseen obstacles.

An additional beneficial component of the project was the research inquiry process. I gained knowledge on how to create precise research questions, carry out literature evaluations, and locate pertinent information sources. I was able to place my thesis in the larger framework of previous research and develop a deeper understanding of the subject matter thanks to this procedure. Additionally, I developed my ability to synthesize data from many sources, which helped me find trends, make connections, and come up with fresh ideas.

The project has been an enriching learning experience, providing me with valuable skills in programming, data analysis, and report writing. It has deepened my understanding of machine learning and artificial intelligence, highlighting both challenges and opportunities in these fields.

Despite the valuable insights gained, I encountered challenges that I was unable to fully overcome. One such challenge was optimizing the performance of certain machine learning models, despite extensive efforts in hyperparameter tuning and feature engineering. Looking back, I realize that I could have dedicated more time to exploring alternative models or ensembles of models, which might have led to better outcomes.

For future iterations of this project, I would approach it differently. I would allocate more time to experimenting with different modeling techniques and thoroughly evaluating their performance. Additionally, I would seek closer collaboration with domain experts to gain a deeper understanding of the clinical context and incorporate their insights into the model development process.

References

David Lapp (1988), 'Public health dataset for heart disease prediction'.
URL: *https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset*

Kumari, M., Godara, S. and Kinoshita, A. (n.d.), 'Comparative study of data mining classification methods in cardiovascular disease prediction', *IJCST* **2**, 304–308.

Pal, M. and Parija, S. (2021), 'Prediction of heart diseases using random forest', *Journal of Physics: Conference Series* **1817**, 012009.

Rankovic, N., Rankovic, D., Lukic, I., Savic, N. and Jovanovic, V. (2023), 'Ensemble model for predicting chronic non-communicable diseases using latin square extraction and fuzzy-artificial neural networks from 2013 to 2019', *Heliyon* **9**(11), e22561.

Shedole, S. S. and Deepika, K. (2016), Predictive analytics to prevent and control chronic disease, *in* 'Proceedings of the Conference Name'.
URL: *https://www.researchgate.net/publication/316530782*

Shetty, A. A. and Naik, C. (2016), 'Different data mining approaches for predicting heart disease', *International Journal of Innovative in Science Engineering and Technology* **5**, 277–281.

Toma, M. and Wei, O. C. (2023), 'Predictive modeling in medicine', *Encyclopedia* **3**(2), 590–601.

Yanwei, X., Wang, J., Zhao, Z. and Gao, Y. (2007), Combination data mining models with new medical data to predict outcome of coronary heart disease, *in* 'Proceedings International Conference on Convergence Information Technology', pp. 868–872.